Chinese Two-part Allegorical Sayings Reading Comprehension: Exploration from Reasoning to Metaphor

Dongyu Su[1, 2, 3, 4], **Yimin Xiao**[4], **Tongguan Wang**[4], **Feiyue Xue**[4], **Junkai Li**[4], **Hui Liu**[4], **Ying Sha**[1, 2, 3, 4*]

[1]Key Laboratory of Smart Farming for Agricultural Animals, Wuhan, China
[2]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan, China
[3]Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan, China
[4]College of Informatics, Huazhong Agricultural University, Wuhan, China
{su_dy, xiaoym, wang_tg, xuefeiyue, junkaili, liuhui_1003}@webmail.hzau.edu.cn, shaying@mail.hzau.edu.cn

Abstract

The **T**wo-**P**art **A**llegorical **S**aying (**TPAS**) is a Chinese linguistic phenomenon with a riddle-explanation structure, and an important component of Chinese metaphors. Existing research has primarily used TPAS to assist other semantic tasks, but lacks in-depth exploration of its intrinsic mechanisms: semantic rhetoric, logical reasoning, and metaphorical expression. To address this gap, we construct the first **C**hinese **TPAS** **R**eading **C**omprehension dataset (**CTRC**), which contains 18,103 TPASs and 75,296 passages. We frame it as a cloze test where the model selects the most suitable TPAS from candidates to fill passage blanks. To tackle the challenges of this CTRC task, we propose a **M**ulti-view **TPAS** **C**ontrastive **L**earning **N**etwork (**MTCLN**). Firstly, the joint vector cross-projection module extracts the rhetorical features of TPAS, such as homophonic puns, through vector space mapping to mitigate the semantic deviations caused by rhetoric. Then, the softened contrastive learning module strengthens the modeling of TPAS logical reasoning through feature association. Finally, the multi-view feature fusion module integrates contextual semantics with diverse TPAS features to facilitate the understanding of metaphorical expressions. Experiments on the CTRC dataset demonstrate that MTCLN achieves an average accuracy of 67.47%, outperforming large language models by 25.48%.

Code — https://github.com/ZhiYue007/TPAS_Data

Introduction

The two-part allegorical saying (TPAS) is a unique Chinese linguistic phenomenon that usually consists of two parts: the riddle and the explanation (Lai 2008). As shown in Figure 1, at the surface semantic level, riddles usually lead to explanations through logical reasoning; at the deep semantic level, TPAS employs rhetoric to bridge surface semantics with metaphors, and the actual meaning of the metaphor often cannot be directly derived from the literal meaning (An and Li 2022; Deng 2023). To accurately understand TPAS, a model needs to comprehend its reasoning logic, the mechanism of deep semantic construction, and relevant Chinese linguistic and cultural knowledge (Gao 2006), which poses a challenge to existing methods including large language models (LLMs). Moreover, accurate comprehension and formal representation of TPAS contribute to downstream tasks, such as text generation (Xu 2024; Cao et al. 2024; Fang et al. 2024), emotional semantic understanding (Su et al. 2024, 2025), machine translation (Liu 2017; Liu et al. 2024), and relevant linguistic phenomena (Hu et al. 2024a, 2025).

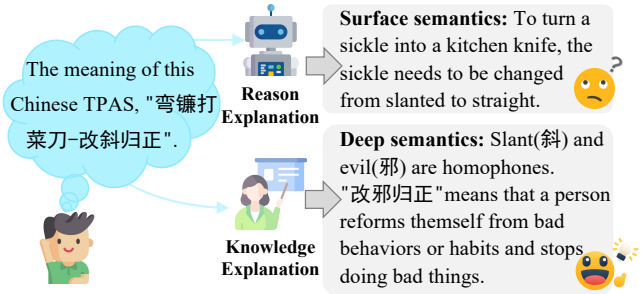


Figure 1: The example demonstrates the process of understanding the surface and deep semantics of TPAS.

In the study of language phenomena, the work related to TPAS mainly utilizes its characteristics to assist in text generation (Xu 2024) and sarcasm detection (Li et al. 2019). To understand specific language phenomena such as idioms and classical Chinese, frameworks like contrastive learning combined with attention have been proposed (Long et al. 2020; Wu et al. 2024a; Wang et al. 2025; Zhang et al. 2024; Xiang et al. 2024). In addition, LLMs such as DeepSeek (Bi et al. 2024; Dai et al. 2024), Hunyuan (Sun et al. 2024), and LLaMA (Zhou et al. 2023; Yang, Cao, and Zhao 2024; Zhao et al. 2025) have achieved significant advancements in metaphor comprehension and semantic analysis, but their exploration of deep-level linguistic phenomena remains limited. Therefore, in-depth investigation into the TPAS comprehension task, which integrates reasoning with multiple rhetorical devices, may offer a new pathway to enhance the semantic comprehension capabilities of existing models.

Wilks' metaphor theory points out that metaphorical semantics change with context (Wilks 1975), which provides a theoretical basis for context-based TPAS semantic under-

*Corresponding author.

Passage& Blanks	他处理每件公务都是#MASK1#——#MASK2#因此得到大家的赞赏。 He handled every official matter #MASK1#——#MASK2#, and was appreciated for it.
Riddle Sets	A.淘米水洗脸(Wash your face with rice water) B.瞎子舔煤球(Blind man licking coal) C.推车下坡(Push a cart downhill) D.红萝卜顶上种小麦(Growing wheat on top of carrots) E.小葱拌豆腐(Scallion mixed with tofu) F.无钱后才断赌(When no money, Stop gambling) G.哑巴吃仙桃(Dumb eating celestial peach)
Explanation Sets	A.粘粘糊糊的(It's sticky) B.无济于事(Be of no avail) C.眼黑嘴也黑(Black eyes and black mouth) D.不图打粮图好看(To look good, not to gain) E.妙不可言(Too wonderful for words) F.一清二白(One cyan, two white) G.使的两股子劲(Make two forces)
Correct TPAS	小葱拌豆腐-一清二白(Mix tofu with scallions, one cyan and two white, Metaphorical meaning: '青' and '清' are homophones. This describes being extremely innocent; It can also be described as very clear)

Figure 2: A sample TPAS cloze test, the riddle is filled in MASK1, and the explanation is filled in MASK2.

standing research. We construct the first Chinese TPAS reading comprehension dataset (CTRC). The data were collected from Zhihu[1], Weibo[2] and TPAS dictionaries, manually annotated with 18,103 TPASs and 75,296 passages, covering the domains of daily conversations, language tests, literary works and news reports. Given the broad application of cloze in education and other fields (Stubbs and Tucker 1974; Stansfield 1980), we borrow ideas from Chinese idiom reading comprehension research (Jiang et al. 2018; Zheng, Huang, and Sun 2019) and transform the TPAS reading comprehension task into a cloze test. Given a sentence with two blanks, the model needs to select the most appropriate options from two candidate sets to fill in the blanks. These two options must together form an accurate TPAS in structure and meaning.

To complete the TPAS cloze test task, the model needs to have the following abilities: (1) Rhetorical recognition: mitigate the semantic changes of TPAS caused by rhetoric such as homophonic puns; (2) Logical reasoning: different riddles have the same explanation and vice versa. The model needs to complete this many-to-many reasoning process; (3) Metaphor comprehension: literal meanings of TPAS are often inconsistent with their actual meanings, and the model needs to obtain the actual semantics.

To address the above challenges, we propose a multi-view TPAS contrastive learning network (**MTCLN**). First, the joint vector cross-projection module extracts the rhetorical features of TPAS, such as homophonic puns, through vector space mapping to mitigate the semantic changes caused by rhetoric. Next, the softened contrastive learning module enhances the understanding of TPAS logical reasoning through feature association. Finally, the multi-view feature fusion module combines contextual semantics with TPAS features to enhance the understanding of metaphors. Experiments on the CTRC dataset show that MTCLN achieves an average accuracy of 67.47%, outperforming LLMs by 25.48%. Our contributions are as follows:

- To our knowledge, we first propose the TPAS reading comprehension task and transform it into a cloze test. We verify its validity in evaluating models' TPAS representation and comprehension, offering a new research sample for related fields.
- We systematically analyze the core challenges of TPAS cloze test and evaluate mainstream LLMs' performance on this task via multiple experiments. The results reveal existing models' shortcomings and provide clear problem orientation for subsequent optimization.
- To address the challenges in TPAS cloze test, we propose a multi-view TPAS contrastive learning network (MTCLN). MTCLN significantly improves the TPAS understanding using joint vector cross-projection, softened contrastive learning, and multi-view feature fusion.

Related Work

Currently, there is relatively little research on TPAS. Therefore, we will introduce two main aspects: the use of TPAS characteristics to assist in the understanding of other tasks and the research related to TPAS reading comprehension.

In the research using TPAS, Li et al. (2019) used SVM to detect TPAS with negative expressions in the sarcasm detection task; Xu (2024) combined TPAS with Chinese crosstalk corpus in the humorous text generation task and used the T5 model combined with the pinyin features of Chinese characters to generate humorous text.

Research on Chinese language phenomena related to TPAS. In the field of idioms, Long et al. (2020) constructed a synonym map by defining the similarity relationship of idioms and combined it with a graph attention network to enhance idiom understanding; Wu et al. (2024a) introduced external knowledge and used contrastive learning to align the literal semantics with the interpretation of idioms and

[1]https://www.zhihu.com/

[2]https://weibo.com/

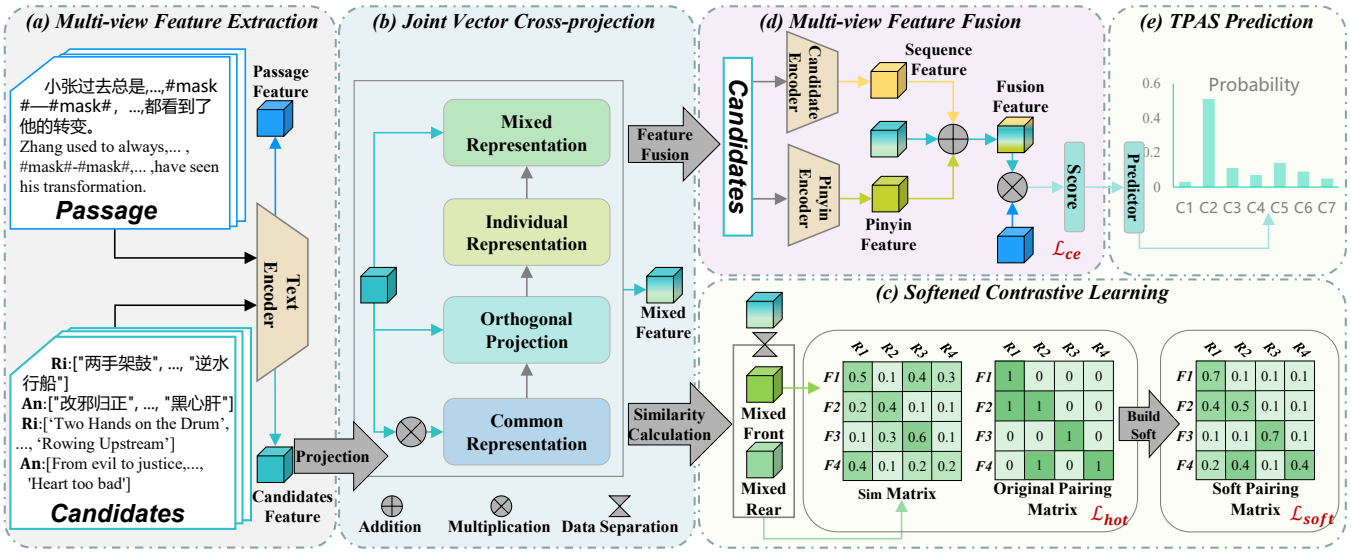


Figure 3: Multi-view TPAS contrastive learning network (MTCLN).

enhanced the contextual understanding through the attention mechanism. In the field of classical Chinese, Zhang et al. (2024) mitigated the negative impact of noisy grammar trees and improved comprehension of ancient Chinese by calculating grammatical feature confidence and combining left and right branch features; Xiang et al. (2024) solved the problem of ephemeral differences by using a parallel corpus of classical and modern Chinese for pre-training, which significantly improved the model's semantic comprehension in classical Chinese.

Task Definition

For t-th input passage $P_t = \{w_1, ..., [\text{MASK}], ..., [\text{MASK}], ..., w_n\}$, where each w_i represents a Chinese character, and [MASK] is marked as two blank parts to be filled. The task is that the model needs to select the most appropriate option from two candidate sets of riddles and explanations $C_i = \{c_1, ..., c_k, ..., c_m\}$ to fill in the blanks, which form a TPAS that is accurate in both structure and meaning. Figure 2 shows an example.

Methodology

We propose the multi-view TPAS contrastive learning network (MTCLN). As shown in Figure 3, the MTCLN consists of five main modules. (1) The multi-view feature extraction module obtains TPAS features through different extractors. (2) The joint vector cross-projection module mitigates the semantic changes caused by rhetoric. (3) The softened contrastive learning module learns the logical reasoning of the TPAS. (4) The multi-view feature fusion module enhances the understanding of TPAS metaphors. (5) The TPAS prediction module computes the best TPAS option.

Multi-view Feature Extraction Module

We first extract TPAS features through different extractors. The RoBERTa model (Cui et al. 2021) extracts semantic features of the paragraph context and TPAS. The [CLS] and [SEP] markers are added at the beginning and end of the passage P_t, respectively, and the positions to be filled are marked with [MASK]. The same processing flow is performed for each riddle or explanation TPAS^*_{k} in the candidate set C_i, where * represents the first or second half of TPAS. As shown in Figure 1, considering the widespread use of rhetoric in TPAS, we leverage linguistic properties of Chinese to aid understanding: first, we obtain the original sequence of TPAS via a Chinese character tokenizer, and then input it into the ChineseBERT model (Sun et al. 2021) to extract Chinese character features TPAS^*_{k} including pinyin and Chinese character shapes. Meanwhile, the sequence tokens of TPAS^*_{k} are inputted into the Embedding layer to obtain the sequence representation:

$$H_t^p = \text{RoBERTa}([\text{CLS}], P_t, [\text{SEP}]), \quad (1)$$

$$H_t^c = \text{RoBERTa}([\text{CLS}], \text{TPAS}^*_{\text{k}}, [\text{SEP}]), \quad (2)$$

$$Y_t^c = \text{ChineseBERT}(\text{TPAS}^*_{\text{k}}), \quad (3)$$

$$E_t^c = \text{Embedding}(\text{TPAS}^*_{\text{k}}), \quad (4)$$

where h_t^p is the feature of the last hidden layer of the [MASK] location as a context-dynamic representation of the TPAS; h_t^c is the [CLS] representation of H_t^c as the original feature of the candidate TPAS. y_t^c is the [CLS] representation of the Chinese character features, and e_t^c is the sequence feature of the TPAS. Multi-view features enable the model to indirectly understand the potential meaning of TPAS in different contexts.

Joint Vector Cross-projection Module

To mitigate semantic deviation from rhetorical disguise in TPAS (where rhetoric-mediated indirect expressions alter actual semantics), we propose a joint vector cross-projection module. Rhetoric modifies TPAS's expressive form but preserves an inherent semantic link between the original and

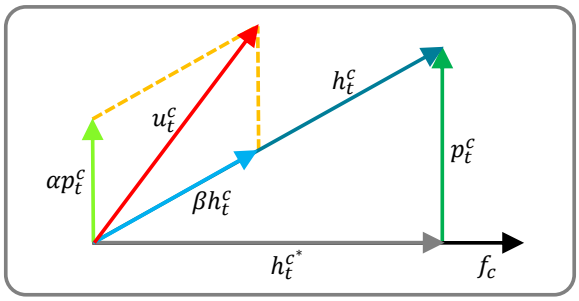


Figure 4: The calculation process of Mixed features.

revised form. In TPAS cloze tasks, each candidate set includes 3 distractors with similar meanings. Leveraging this, TPAS semantics can be split into common and personalized components: common semantics are extracted from the candidate set, while rhetorical changes are attributed to personalized semantics. Adjusting the weight of personalized semantics mitigates the deviation.

The module initially integrates the original features of the candidate TPASs to extract the common features of the semantics. Subsequently, the original features of the candidates are orthogonally projected onto the common feature space to isolate the individual semantic components. Finally, the original features and individual features are combined to generate Mixed features that extract the individual semantic. As shown in Figure 4, the module obtains common features f_c by candidate original features h_t^c as follows:

$$f_c = \frac{\sum_{t=1}^{m} d_t h_t^c}{\sum_{t=1}^{m} d_t}, \tag{5}$$

where $d_i \in \mathbb{R}$ is the weight of the common feature. The original feature h_t^c is projected onto the common feature f_c to obtain the projected feature $h_t^{c^*}$, and the individual feature p_t^c is extracted by calculating the difference from h_t^c:

$$h_t^{c^*} = \frac{h_t^c \cdot f_c}{|f_c|} \frac{f_c}{|f_c|}, \tag{6}$$

$$p_t^c = h_t^c - h_t^{c^*}. \tag{7}$$

The final TPAS needs to adjust the individual semantics and retain the original features, therefore a weighted approach is used to generate Mixed features u_t^c:

$$u_t^c = \alpha p_t^c + \beta h_t^c, \tag{8}$$

where α and β are the weights of individual and original features, respectively, and satisfy the sum of α, β is 1.

Softened Contrastive Learning Module

In the contextual semantic reasoning of TPAS, the model must complete the text blanks while ensuring the consistency of reasoning between the riddle and the explanation. Since there is a leap in the semantics between the two parts of TPAS, we refer to the work of Hu et al. (2024b) and Wu et al. (2024b) to enhance their semantic associations through contrastive learning. However, traditional contrastive learning methods enhance the similarity through one-hot encoding but ignore the relationship between non-matching TPASs, which cannot satisfy the TPAS task for many-to-many relationships. To this end, we introduce a softened contrastive learning module that shifts partial confidence from positive to negative samples through a label smoothing strategy, allows for weak similarity between negative samples, and uses a data matrix of many-to-many pairs. This approach enhances sample correlation and improves the model's logical reasoning.

The TPAS joint feature u_t^c of the t-th candidate group is first obtained and separated into two parts representing the riddle and the explanation, denoted as $[u_t^f, u_t^r]_{i=1}^M$. Next, the normalized text similarity function is input to calculate the similarity value for each pair of TPAS:

$$S_{ij} = \frac{\exp(\text{sim}(u_t^f, u_t^r)/\tau)}{\sum_{j=1}^{n} \exp(\text{sim}(u_t^f, u_t^r)/\tau)}, \tag{9}$$

where sim is the cosine similarity function. τ is a learnable temperature hyperparameter with an initial value of 0.07. We construct a multi-paired data matrix through the data dictionary. The labels of the i-th pair are denoted as $y_i = \{y_{ij}\}_{j=1}^M$, where y_{ij} is 1 for paired data and 0 for unpaired. The data labels are adjusted to the following form:

$$\tilde{\mathbf{y}} = \begin{cases} \frac{1}{M-1} \max(\text{sim}(u_t^f, u_t^r) - d, 0), & \text{if } y = 0, \\ \max(\text{sim}(u_t^f, u_t^r), 1 - d), & \text{if } y = 1, \end{cases} \tag{10}$$

where the data smoothing value d is 0.2, and 1 represents a full 1 matrix. Since the matrix similarity matching mode is adopted, the cross-entropy loss function is replaced with the Kullback-Leibler divergence function:

$$\mathcal{L}_{soft} = \frac{1}{M} \sum_{i=1}^{M} \text{KL}(\tilde{y}_i || s_i). \tag{11}$$

The clip loss function incorporates the auxiliary loss $\mathcal{L}_{hot}$ generated by the original data matrix with unmodified labels to simultaneously learn the many-to-many logical reasoning of the TPAS and maintain the original paired understanding, with the clip loss $\mathcal{L}_{clip}$ as follows:

$$\mathcal{L}_{hot} = \frac{1}{M} \sum_{i=1}^{M} \text{KL}(y_i || s_i), \tag{12}$$

$$\mathcal{L}_{clip} = \mathcal{L}_{soft} + \mathcal{L}_{hot}. \tag{13}$$

Multi-view Feature Fusion Module

The joint vector cross-projection and softened contrastive learning module enhances TPAS understanding through candidate differences and associations, but it is still limited in contextual metaphor understanding. We propose a multi-view feature fusion module to resolve literal-actual meaning inconsistency in TPAS, extracting optimal TPAS hybrid features and fusing them with contextual features for enhanced multi-scene metaphor understanding.

To make the final feature more effective in understanding the TPAS rhetorical and metaphorical characteristics, the optimal TPAS hybrid feature q_t^c is obtained by taking the joint

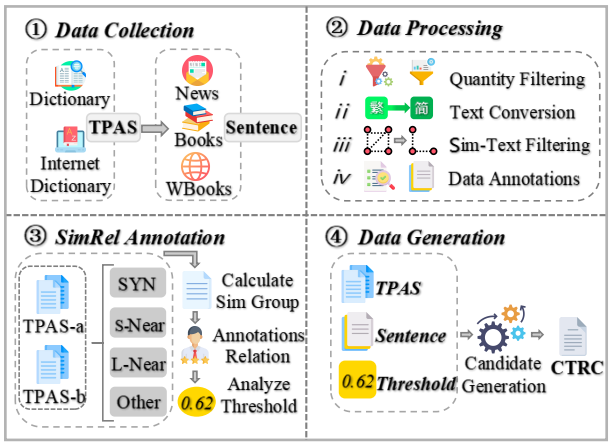


Figure 5: The construction process of the CTRC dataset, where the four relations defined by the similarity relation annotation are: Synonyms, Semantic Near-synonyms, Literal Near-synonyms, and Other words.

feature u_t^c as the dominant feature and combining it with the Chinese character features y_t^c and the sequence feature e_t^c for fusion, with the following equation:

$$\begin{cases} \mu_y^c = W_{y_t} y_t^c + b_{y_t}, \\ \mu_e^c = W_{e_t} e_t^c + b_{e_t}, \end{cases} \tag{14}$$

$$q_t^c = u_t^c + \lambda \mu_y^c + \gamma \mu_e^c, \tag{15}$$

where λ and γ are different parameter weights. To enhance TPAS's multi-scene metaphor understanding, hybrid features q_t^c are used to score the contextual features h_t^p for prediction, and scoring results serve as final information:

$$\mathrm{P}(c_i|q_t^c, h_t^p) = \frac{\exp(w \cdot (q_t^c \otimes h_t^p) + b)}{\sum_{k=1}^{M} \exp(w \cdot (q_t^c \otimes h_t^p) + b)}, \tag{16}$$

where $w \in R^d$ denotes the model parameters, b is the bias term, and $\otimes$ denotes the elemental multiplication.

TPAS Prediction Module

TPAS's prediction loss $\mathcal{L}_{ce}$ is obtained by minimizing the cross-entropy loss between predicted and true values:

$$\mathcal{L}_{ce} = -\sum_{j=1}^{m} o_j \log \mathrm{P}(c_j|q_t^c, h_t^p). \tag{17}$$

For a candidate set containing m riddles and explanations, o_j is the one-hot encoded vector corresponding to the target label. The final training loss $\mathcal{L}_{ft}$ is the sum of the prediction loss $\mathcal{L}_{ce}$ and the clip loss $\mathcal{L}_{clip}$ with the following equation:

$$\mathcal{L}_{ft} = \mathcal{L}_{clip} + \mathcal{L}_{ce}. \tag{18}$$

Experiments

Evaluation Setup

Datasets: Figure 5 outlines the construction process of the CTRC dataset, which is divided into four stages: data collection, data processing, similarity relation annotation and data generation. The CTRC is divided into training, development, and testing sets in the ratio of 7:1:2. Each sample contains one passage and two candidate sets, which consist of the target, three approximations, and three irrelevant TPAS. To evaluate the robustness of the model, we reconstruct the candidate composition of the test set to generate three generalization sets: the Asy set is used to verify the reasoning ability, and its candidate set only pairs the target TPAS; the Sim set is used to verify the rhetorical and metaphorical understanding ability, and its candidate set is composed of high similarity TPASs; and the Ran set is used to verify the semantic understanding ability, and its candidate set is composed of irrelevant TPASs. The CTRC dataset contains 18,103 TPASs and 75,296 passages, and the distribution is shown in Table 1.

	Train	Dev	Test	Asy	Sim	Ran
Passage	51,702	9,797	13,797	13,797	13,797	13,797
TPAS	18,103	16,510	17,953	10,341	16,648	17,981

Table 1: Distribution of 5 sub-datasets in CTRC dataset.

Baselines: Our approach is compared with three traditional text models (Zheng, Huang, and Sun 2019), three BERT-based Chinese pre-training models (Cui et al. 2021), ChineseBERT (Sun et al. 2021) incorporating Chinese character features, and Chinese Mengzi (Zhang et al. 2021) with knowledge distillation. For LLMs, we select the five models with the best reasoning and knowledge performance in OpenCompass[3] Top 10, and test the few-shot scheme based on GPT-4o-20241120 and the LoRA fine-tuning method of DeepSeek-R1-1.5B.

Implement Details: For model training, we set the maximum sequence length to 128, the initial learning rate to 5e-5, and the batch size to 32. The optimizer uses AdamW with a warm-up linear scheduler. Training runs for 50 epochs with an early stop after 8 consecutive invalid iterations. Experiments are conducted on an NVIDIA A100 40G, using PyTorch 2.1.2 and Transformers 4.41.2.

Accuracy metrics: We use accuracy as the main evaluation metric to measure model performance. Due to the specificity of the task, a sample is considered to be correctly filled only when both empty spaces in the sample are correctly filled. The accuracy rate is calculated as the ratio of the number of correctly filled samples to the total number of samples. In addition, we calculated the mean value of the accuracy rate to assess the overall performance of the model fully.

Results Analysis

Comparison with Baselines: Table 2 shows each model's performance under optimal parameters. MTCLN performs best on the main test task and the three auxiliary tasks. Notably, all model accuracy on the Sim set is lower than other sets, except for MTCLN, indicating that rhetorical and metaphorical understanding is an important challenge for TPAS. Mengzi performs best in comparison with other Base models, probably because it uses a large corpus of Chinese

[3]https://rank.opencompass.org.cn/home

Method	Dev-Acc	Test-Acc	Asy-Acc	Sim-Acc	Ran-Acc	Avg-Acc
Human	-	85.25	81.50	73.25	89.50	82.37
LM(Zheng, Huang, and Sun 2019)	36.60	36.82	30.35	28.26	38.13	34.03
AR(Zheng, Huang, and Sun 2019)	47.40	48.38	39.31	36.24	49.91	44.24
SAR(Zheng, Huang, and Sun 2019)	43.34	44.46	36.09	34.29	46.29	40.89
BERT-WWM(Cui et al. 2021)	51.64	51.17	43.08	41.97	52.23	48.01
RoBERTa(Cui et al. 2021)	56.45	55.87	47.98	45.98	57.26	52.70
macBERT(Cui et al. 2021)	56.03	56.56	48.08	45.80	56.88	52.67
ChineseBERT(Sun et al. 2021)	54.98	55.71	46.35	44.82	56.60	51.69
MengZi(Zhang et al. 2021)	57.17	57.91	49.09	46.46	58.63	53.85
Doubao-pro-32k	32.20	36.61	29.11	25.88	40.73	32.90
Claude 3.5 Sonnet	42.05	43.08	36.20	28.08	48.97	39.67
GLM-4-Plus	39.26	39.26	33.52	28.23	45.00	37.05
Qwen2.5-72B-Instruct	38.38	39.70	34.85	27.35	46.02	37.26
GPT-4o-20241120	44.11	42.94	34.35	28.32	49.70	39.88
GPT-4o-1-shot	46.91	43.52	36.47	28.08	54.11	41.81
GPT-4o-3-shot	46.76	46.32	36.91	27.79	52.20	41.99
GPT-4o-5-shot	46.61	44.85	35.44	28.52	52.50	41.58
DeepSeek-R1-1.5B	43.41	43.19	31.09	25.38	48.14	38.24
MTCLN (Ours)	**72.85**	**73.13**	**57.22**	**58.34**	**75.81**	**67.47**

Table 2: Comparison results (%) with baseline models on the CTRC dataset.

language phenomena in its pre-training. In addition, the accuracy of MTCLN on the Asy set is lower than on the Sim set, possibly due to the closer TPAS feature space after training, making the TPAS of many-to-many relations difficult to distinguish, as shown in Figure 7. Therefore, it is more challenging to simultaneously process reasoning and mitigate approximate differences in TPAS understanding tasks.

Comparison with LLMs: As shown in Table 2, MTCLN is still competitive in performance compared to LLMs. Further analysis found that: (1) LLMs performed better on the Asy set than on the Sim set, indicating that they are adept at reasoning but have limitations in understanding Chinese metaphors; (2) The effect of few-shot learning does not continue to improve with the increase in the number of prompt samples, which may be because metaphor understanding requires support from a specific cultural background; (3) Although DeepSeek-R1-1.5B has a large number of parameters, the effect is poor after Lora fine-tuning, indicating that metaphor understanding cannot rely solely on the number of model parameters.

Comparison with Human Evaluation: To evaluate the difference between the model and human performance, we randomly select 200 samples from each dataset to form the evaluation set and invited four native Chinese university students to participate in the test. Table 2 shows that participants outperform the model on average, with the worst performance on the Sim set, highlighting metaphor understanding as a key TPAS challenge. The maximum human-model performance gap on the Asy set is 24.28%. To investigate, we had participants identify 100 sampled TPASs. Though familiar with less than 50% of them, participants still derived correct explanations via knowledge correlation and reasoning—showing humans outperform the model in logical reasoning, a key for model improvement.

Method	Dev	Test	Asy	Sim	Ran	Avg
w/o JP	72.70	72.89	56.54	57.41	75.28	66.96
w/o SL	64.29	69.15	56.19	46.51	67.77	59.18
w/o MF	72.17	72.79	56.66	57.10	75.22	66.78
w/o SL+MF	63.54	63.79	52.67	46.00	67.35	58.67
w/o JP+MF	58.05	58.67	50.04	47.51	59.07	54.66
w/o JP+SL	63.74	63.68	53.24	46.21	67.23	58.82
w/o ALL	56.45	55.87	47.98	45.98	57.26	52.70
MTCLN	**72.85**	**73.13**	**57.22**	**58.34**	**75.81**	**67.47**

Table 3: MTCLN ablation study on the CTRC dataset.

Ablation Study

We evaluate the contribution of each module in MTCLN through ablation experiments. As shown in Table 3, joint vector cross-projection (JP), softened contrast learning (SL), and multi-view feature fusion (MF) are sequentially eliminated from MTCLN. When JP or MF is removed, there is a slight decrease in accuracy, suggesting that capturing TPAS differences or combining contextual semantics can improve model performance. Accuracy decreases when removing SL, indicating that model enhancement through feature processing and fusion is limited, while semantic association can effectively improve performance. When removing JP+MF, the accuracy decreases the most, indicating that semantic association significantly improves the model's understanding of TPAS. Removing JP+SL or SL+MF also leads to a significant decrease in accuracy. These results demonstrate that the concerted work of the MTCLN modules is crucial for completing the TPAS cloze test task.

Hyperparametric Analysis

We perform a detailed hyperparameter optimization of the parameters α and β in the joint vector cross-projection

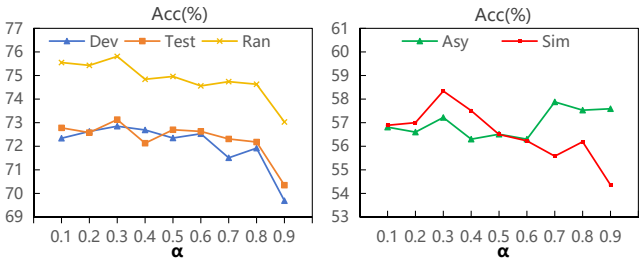
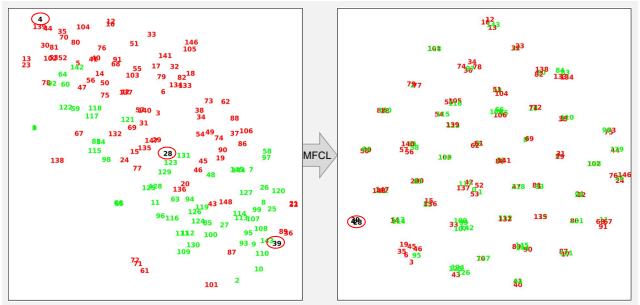


Figure 6: The impact of individual representation weights α.



(a) Representation distribution of 100 TPASs before model training.

(b) Representation distribution of 100 TPASs after model training.

Figure 7: The distribution of TPAS representations changed significantly before and after model training. The green color in the figure indicates riddles, red indicates explanations, and the black color indicates the two specially marked TPAS. For example, "冷水烫鸡" (riddle, #28: Chicken scalded with cold water), "名牌牙刷" (riddle, #39: Famous toothbrush) and "一毛不拔" (explanation, #4: Can't lose a hair, actually describes a person who is very stingy) constitute two pairs of TPAS respectively and has the same more profound meaning. After training, the representation distance of these TPAS pairs is shortened, indicating that the reasoning ability of the model is improved.

model. As shown in Figure 6, the model achieves optimal performance when α takes the value of 0.3, and β takes the value of 0.7. This result shows that in the rhetorical challenge of TPAS, the local semantic effect of individual features combined with the main semantics of common features can improve the overall semantic understanding and mitigate the semantic differences caused by rhetoric.

Visualization

To verify the model's reliability, we extract 100 TPAS samples for analysis, of which 60% are manually selected samples with many-to-many complex relationships, and the remaining 40% are random. As shown in Figure 7, we project the features before and after training into 2D space for observation using the t-SNE method (Van der Maaten and Hinton 2008). Cosine similarity results show that the overall semantic similarity aggregation degree increased by 8.14%, and the similarity before and after TPAS increased by 6.34%. After training, the TPAS feature representation is significantly converged, which indicates that the model's logical reasoning ability is effectively improved.

Case Study

To illustrate how MTCLN works and to compare the performance of other models, we provide one correct case and three types of error cases. Given that GPT-4 performs best in the test, we use it as a proxy for LLMs to explore their decision-making rationale.

Correct example: GPT-4 grasps the central meaning of passages by effectively interpreting metaphors and establishing relevant knowledge associations. For example, the metaphorical meaning of the idiom "爱不释手"(I love it so much that I can't bear to let go) is closely related to its literal interpretation, which facilitates comprehension by both GPT-4 and MTCLN.

Type I error: The model is unable to understand the semantic changes induced by rhetoric. GTP-4 fails to recognize the homophonic pun between '舅'(uncle) and '旧'(old). This failure results in an inability to capture the semantic change induced by rhetoric. In contrast, MTCLN effectively identifies the rhetorical differences and successfully understands the intended meaning by highlighting these distinctions.

Type II error: The model exhibits deficiencies in performing logical reasoning for TPAS. Although the options of GPT-4 "脚踩棉花堆–拖拖拉拉"(step on the cotton pile–propagate) are consistent with the contextual semantics, there is no logical reasoning process for the two options. The MTCLN improves logical reasoning through semantic associations, leading to more appropriate choices.

Type III error: The model fails to capture the explanation's metaphorical expression. The model needs to identify the difference between the literal and metaphorical meanings of the idiom "放任自流"(The river flows on its own, which actually means letting it develop without intervention). Idioms are an important part of Chinese metaphors and frequently appear in TPAS explanations. Since metaphor comprehension relies on deep knowledge of the Chinese context, models such as GPT-4 are still insufficient in understanding and applying idioms.

Conclusion

We first construct a Chinese TPAS reading comprehension dataset (CTRC). To address the challenges of rhetorical recognition, logical reasoning, and Metaphor comprehension in TPAS, we propose a multi-view TPAS contrastive learning network (MTCLN). First, the joint vector cross-projection module extracts the rhetorical features of TPAS, such as homophonic puns, through vector space mapping to mitigate the semantic changes caused by rhetoric. Then, the softened contrastive learning module enhances the understanding of TPAS logical reasoning through feature association. Finally, the multi-view feature fusion module combines contextual semantics with TPAS features to enhance the understanding of metaphors. Experiments on the CTRC dataset show that the MTCLN achieves an average accuracy of 67.47%, outperforming LLMs by 25.48%.

Ethical Statement

This study strictly follows the data usage protocols of public online social platforms. The TPAS data are mainly obtained from the Chinese Xinhua Dictionary database and the online TPAS Dictionary website. Passage samples are mainly from Zhihu, Weibo, and the TPAS dictionary. All information is used for scientific research only.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62272188). Thanks to the anonymous reviewers for their hard efforts!

References

An, S.; and Li, S. 2022. Two-part Allegorical Saying of Insect in Northeast Dialect and Its Cultural Implication. In *2021 International Conference on Social Development and Media Communication (SDMC 2021)*, 1473–1476. Atlantis Press.

Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Cao, J.; Liu, Y.; Shi, Y.; Ding, K.; and Jin, L. 2024. WenMind: A comprehensive benchmark for evaluating large language models in Chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37: 51358–51410.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1280–1297.

Deng, L. 2023. Study on the Origin and Evolution of Xiehouyu. *Journal of Education and Educational Research*, 4(1): 138–139.

Fang, J.; Lu, T.; Yao, Y.; Jiang, Z.; Xu, X.; Zhang, N.; and Chen, H. 2024. CKnowEdit: A New Chinese Knowledge Editing Dataset for Linguistics, Facts, and Logic Error Correction in LLMs. *arXiv e-prints*, arXiv–2409.

Gao, L. 2006. Language contact and convergence in computer-mediated communication. *World Englishes*, 25(2): 299–308.

Hu, Y.; Li, J.; Chen, M.; Su, D.; Wang, T.; and Sha, Y. 2025. Keyword-Oriented Multimodal Modeling for Euphemism Identification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Hu, Y.; Li, J.; Wang, T.; Su, D.; Su, G.; and Sha, Y. 2024a. A Unified Generative Framework for Bilingual Euphemism Detection and Identification. In *Findings of the Association for Computational Linguistics ACL 2024*, 6753–6766.

Hu, Y.; Li, J.; Wu, M.; Huang, Z.; Chen, G.; and Sha, Y. 2024b. Uncovering and Mitigating the Hidden Chasm: A Study on the Text-Text Domain Gap in Euphemism Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, 18270–18278.

Jiang, Z.; Zhang, B.; Huang, L.; and Ji, H. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 154–158.

Lai, H.-l. 2008. Understanding and classifying two-part allegorical sayings: Metonymy, metaphor, and cultural constraints. *Journal of Pragmatics*, 40(3): 454–474.

Li, A.-R.; Chersoni, E.; Xiang, R.; Huang, C.-R.; and Lu, Q. 2019. On the "easy" task of evaluating Chinese irony detection. In *Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation*, 452–460. Waseda University.

Liu, C.; Koto, F.; Baldwin, T.; and Gurevych, I. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2016–2039.

Liu, J. 2017. A Study of the Translation of Two-Part Allegorical Sayings in Hongloumeng. In *Asia International Symposium on Language, Literature and Translation*, 68.

Long, S.; Wang, R.; Tao, K.; Zeng, J.; and Dai, X. 2020. Synonym Knowledge Enhanced Reader for Chinese Idiom Reading Comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3684–3695.

Stansfield, C. 1980. The cloze procedure as a progress test. *Hispania*, 63(4): 715–718.

Stubbs, J. B.; and Tucker, G. R. 1974. The cloze test as a measure of English proficiency. *The Modern Language Journal*, 58(5/6): 239–241.

Su, G.; Wu, M.; Huang, Z.; Zhang, Y.; Wang, T.; Hu, Y.; and Sha, Y. 2024. Refine, align, and aggregate: multi-view linguistic features enhancement for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, 3212–3228.

Su, G.; Zhang, Y.; Wang, T.; Wu, M.; and Sha, Y. 2025. Unified Grid Tagging Scheme for Aspect Sentiment Quad Prediction. In *Proceedings of the 31st International Conference on Computational Linguistics*, 3997–4010.

Sun, X.; Chen, Y.; Huang, Y.; Xie, R.; Zhu, J.; Zhang, K.; Li, S.; Yang, Z.; Han, J.; Shu, X.; et al. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Sun, Z.; Li, X.; Sun, X.; Meng, Y.; Ao, X.; He, Q.; Wu, F.; and Li, J. 2021. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2065–2075.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, T.; Wu, M.; Su, G.; Su, D.; Hu, Y.; Huang, Z.; and Sha, Y. 2025. MChIRC: A Multimodal Benchmark for Chinese Idiom Reading Comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 24, 25398–25406.

Wilks, Y. 1975. A preferential, pattern-seeking, semantics for natural language inference. *Artificial intelligence*, 6(1): 53–74.

Wu, M.; Hu, Y.; Zhang, Y.; Zhi, Z.; Su, G.; and Sha, Y. 2024a. Mitigating idiom inconsistency: A multi-Semantic Contrastive Learning Method for Chinese idiom reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17, 19243–19251.

Wu, M.; Su, G.; Zhang, Y.; Huang, Z.; and Sha, Y. 2024b. Refining Idioms Semantics Comprehension via Contrastive Learning and Cross-Attention. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13785–13795.

Xiang, J.; Liu, M.; Li, Q.; Qiu, C.; and Hu, H. 2024. A cross-guidance cross-lingual model on generated parallel corpus for classical Chinese machine reading comprehension. *Information Processing & Management*, 61(2): 103607.

Xu, R. 2024. Exploring Chinese Humor Generation: A Study on Two-part Allegorical Sayings. In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Yang, Y.; Cao, Z.; and Zhao, H. 2024. LaCo: Large Language Model Pruning via Layer Collapse. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 6401–6417.

Zhang, S.; Wang, P.; Li, Z.; Hou, J.; and Hu, Q. 2024. Confidence-based Syntax encoding network for better ancient Chinese understanding. *Information Processing & Management*, 61(3): 103616.

Zhang, Z.; Zhang, H.; Chen, K.; Guo, Y.; Hua, J.; Wang, Y.; and Zhou, M. 2021. Mengzi: Towards lightweight yet ingenious pre-trained models for chinese. *arXiv preprint arXiv:2110.06696*.

Zhao, S.; Zhou, Y.; Ren, Y.; Chen, Z.; Jia, C.; Zhe, F.; Long, Z.; Liu, S.; and Lan, M. 2025. F\ux\i: A Benchmark for Evaluating Language Models on Ancient Chinese Text Understanding and Generation. *arXiv preprint arXiv:2503.15837*.

Zheng, C.; Huang, M.; and Sun, A. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 778–787.

Zhou, K.; Zhu, Y.; Chen, Z.; Chen, W.; Zhao, W. X.; Chen, X.; Lin, Y.; Wen, J.-R.; and Han, J. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.